BAR CODE LABEL	U.S. PATENT APPLICATION		

SERIAL NUMBER	FILING DATE	CLASS	GROUP ART UNIT
08/212,433	03/14/94	435	1807

APPLICANT

JOHN R. YATES III, SEATTLE, WA; JAMES K. ENG, SEATAC, WA.

CONTINUING DATA********************
VERIFIED

———

REC'D 2 0 APR 1995
WIPO PCT

FOREIGN/PCT APPLICATIONS***********
VERIFIED

———

PRIORITY DOCUMENT

FOREIGN FILING LICENSE GRANTED 05/06/94   ***** SMALL ENTITY *****

STATE OR COUNTRY	SHEETS DRAWING	TOTAL CLAIMS	INDEPENDENT CLAIMS	FILING FEE RECEIVED	ATTORNEY DOCKET NO.
WA	10	21	3	$366.00	163362

ADDRESS

RICHARD L. HUGHES
TOWNSEND AND TOWNSEND KHOURIE AND CREW
STEUART STREET TOWER
ONE MARKET PLAZA, 20TH FLOOR
SAN FRANCISCO, CA 94105

TITLE

USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO
IDENTIFY AMINO ACID SEQUENCES IN DATABASES

This is to certify that annexed hereto is a true copy from the records of the United States
Patent and Trademark Office of the application which is identified above.

By authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

MAR 2 4 1995

Date   Certifying Officer

08/212433

PATENT APPLICATION SERIAL NO._____

U.S. DEPARTMENT OF COMMERCE
PATENT AND TRADEMARK OFFICE
FEE RECORD SHEET

BF10076 03/22/94 08212433   20-1430 010 201   366.00CH 16336-2

PTO-1556
(5/87)

USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO IDENTIFY AMINO ACID SEQUENCES IN DATABASES

The present invention is directed to mass
spectrometry of peptides and, in particular, to correlating
fragmentation patterns of peptide fragments obtained from mass
spectrometry with amino acid sequences stored in a database.

BACKGROUND OF THE INVENTION

A number of approaches have been used in the past
for applying the analytic power of mass spectrometry to
peptides. Tandem mass spectrometry (MS/MS) techniques have
been particularly useful. In tandem mass spectrometry, the
peptide or other input (commonly obtained from a
chromatography device) is applied to a first mass spectrometer
which serves to select, from a mixture of peptides, a target
peptide of a particular mass or molecular weight. The target
peptide is then activated or fragmented to produce a mixture
of the "target" or parent peptide and various component
fragments, typically peptides of smaller mass. This mixture
is then applied to a second mass spectrometer which generates
a fragment spectrum. This fragment spectrum will typically be
expressed in the form of a bar graph having a plurality of
peaks, each peak indicating the mass-to-change ratio (m/z) of
a detected fragment and having an intensity value.

Although the bare fragment spectrum can be of some
interest, it is often desired to use the fragment spectrum to
identify the peptide (or the parent protein) which resulted in
the fragment mixture. Previous approaches have typically
involved using the fragment spectrum as a basis for
hypothesizing one or more candidate amino acid sequences.
This procedure has typically involved human analysis by a
skilled researcher, although at least one automated procedure
has been described. John Yates, III, et al., "Computer Aided
Interpretation of Low Energy MS/MS Mass Spectra of Peptides"

2

Techniques In Protein Chemistry II (1991), pp. 477-485,
incorporated herein by reference. The candidate sequences can
then be compared with known amino acid sequences of various
proteins in the protein sequence libraries.

5   The procedure which involves hypothesizing candidate
amino acid sequences based on fragment spectra is useful in a
number of contexts but also has certain difficulties.
Interpretation of the fragment spectra so as to produce
candidate amino acid sequences is time-consuming, often
10   inaccurate, highly technical and in general can be performed
only by a few laboratories with extensive experience in tandem
mass spectrometry. Reliance on human interpretation often
means that analysis is relatively slow and lacks strict
objectivity. Approaches based on peptide mass mapping are
15   limited to peptide masses derived from an intact homogenous
protein generated by specific and known proteolytic cleavage
and thus are not generally applicable to mixtures of proteins.
         Accordingly, it would be useful to provide a system
for correlating fragment spectra with known protein sequences
20   while avoiding the delay and/or subjectivity involved in
hypothesizing or deducing candidate amino acid sequences from
the fragment spectra.

SUMMARY OF THE INVENTION
25   According to the present invention, known amino acid
sequences, e.g., in a protein sequence library, are used to
calculate or predict one or more candidate fragment spectra.
The predicted fragment spectra are then compared with an
experimentally-derived fragment spectrum to determine the best
30   match or matches. Preferably, the parent peptide, from which
the fragment spectrum was derived has a known mass. Sub-
sequences of the various sequences in the protein sequence
library are analyzed to identify those sub-sequences
corresponding to a peptide whose mass is equal to (or within a
35   given tolerance of) the mass of the parent peptide in the
fragment spectrum. For each sub-sequence having the proper
mass, a predicted fragment spectrum can be calculated, e.g.,
by calculating masses of various amino acid subsets of the

candidate peptide. The result will be a plurality of
candidate peptides, each with a predicted fragment spectrum.
The predicted fragment spectra can then be compared with the
fragment spectrum derived from the tandem mass spectrometer to

5   identify one or more proteins having sub-sequences which are
likely to be identical with the sequence of the peptide which
resulted in the experimentally-derived fragment spectrum.

BRIEF DESCRIPTION OF THE DRAWINGS

10   Fig. 1 is a block diagram depicting previous methods
for correlating tandem mass spectrometer data with sequences
from a protein sequence library;

        Fig. 2 is a block diagram showing a method for
correlating tandem mass spectrometer data with sequences from

15   a protein sequence library according to an embodiment of the
present invention;

        Fig. 3 is a flow chart showing steps for correlating
tandem mass spectrometry data with amino acid sequences,
according to an embodiment of the present invention;

20   Fig. 4 is a flow diagram showing details of a method
for the step of identifying candidate sub-sequences of Fig. 3;

        Fig. 5 is a fragment mass spectrum for a peptide of
a type that can be used in connection with the present
invention; and

25   Figs. 6A-6D are flow charts showing an analysis
method, according to an embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

        Before describing the embodiments of the present

30   invention, it will be useful to describe, in greater detail, a
previous method. As depicted in Fig. 1, the previous method
is used for analysis of an unknown peptide 12. Typically the
peptide will be output from a chromatography column which has
been used to separate a partially fractionated protein. The

35   protein can be fractionated by, for example, gel filtration
chromatography and/or high performance liquid chromatography
(HPLC). The sample 12 is introduced to a tandem mass
spectrometer 14 through an ionization method such as

4

electrospray ionization (ES). In the first mass spectrometer, a peptide ion is selected, so that a targeted component of a specific mass, is separated from the rest of the sample 14a. The targeted component is then activated or decomposed. In

5   the case of a peptide, the result will be a mixture of the ionized parent peptide ("precursor ion") and component peptides of lower mass which are ionized to various states. A number of activation methods can be used including collisions with neutral gases (also referred to as collision induced

10   dissolution). The parent peptide and its fragments are then provided to the second mass spectrometer 14c, which outputs an intensity and m/z for each of the plurality of fragments in the fragment mixture. This information can be output as a fragment mass spectrum 16. Fig. 5 provides an example of such

15   a spectrum 16. In the spectrum 16 each fragment ion is represented as a bar graph whose abscissa value indicates the mass-to-charge ratio (m/z) and whose ordinate value represents intensity. According to previous methods, in order to correlate a fragment spectrum with sequences from a protein

20   sequence library, a fragment sequence was converted into one or more amino acid sequences judged to correspond to the fragment spectrum. In one strategy, the weight of each of the amino acids is subtracted from the molecular weight of the parent ion to determine what might be the molecular weight of

25   a fragment assuming, respectively, each amino acid is in the terminal position. It is determined if this fragment mass is found in the actual measured fragment spectrum. Scores are generated for each of the amino acids and the scores are sorted to generate a list of partial sequences for the next

30   subtraction cycle. Cycles continue until subtraction of the mass of an amino acid leaves a difference of less than 0.5 and greater than -0.5. The result is one or more candidate amino acid sequences 18. This procedure can be automated as described, for example, in Yates III (1991) supra. One or

35   more of the highest-scoring candidate sequences can then be compared 21 to sequences in a protein sequence library 20 to try to identify a protein having a sub-sequence similar or

identical to the sequence believed to correspond to the
peptide which generated the fragment spectrum 16.

    Fig. 2 shows an overview of a process according to
the present invention. According to the process of Fig. 2, a
5   fragment spectrum 16 is obtained in a manner similar to that
described above for the fragment spectrum depicted in Fig. 1.
Specifically, the sample 12 is provided to a tandem mass
spectrometer 14. Procedures described below use a two-step
process to acquire ms/ms data. However the present invention
10   can also be used in connection with mass spectrometry
approaches currently being developed which incorporate
acquisition of ms/ms data with a single step. In one
embodiment ms/ms spectra would be acquired at each mass. The
first ms would separate the ions by mass-to-charge and the
15   second would record the ms/ms spectrum. The second stage of
ms/ms would acquire, e.g. 5 to 10 spectra at each mass
transformed by the first ms.

    A number of mass spectrometers can be used including
a triple-quadruple mass spectrometer, a Fourier-transform
20   cyclotron resonance mass spectrometer, a tandem time-of-flight
mass spectrometer and a quadrupole ion trap mass spectrometer.
In the process of Fig. 2, however, it is not necessary to use
the fragment spectrum as a basis for hypothesizing one or more
amino acid sequences. In the process of Fig. 2, sub-sequences
25   contained in the protein sequence library 20 are used as a
basis for predicting a plurality of mass spectra 22, e.g.,
using prediction techniques described more fully below.

    A number of sequence libraries can be used,
including, for example, the Genpept database, the GenBank
30   database (described in Burks, et al., "GenBank: Current status
and future directions in *Methods in Enzymology*", 183:3
(1990)), EMBL data library (described in Kahn, et al., "EMBL
Data Library," Methods in Enzymology, 183:23 (1990)), the
Protein Sequence Database (described in Barker, et al.,
35   "Protein Sequence Database," Methods in Enzymology, 1983:31
(1990), SWISS-PROT (described in Bairoch, et al., "The SWISS-
PROT protein sequence data bank, recent developments," Nucleic
Acids Res., 21:3093-3096 (1993)), and PIR-International

(described in "Index of the Protein Sequence Database of the
International Association of Protein Sequence Databanks (PIR-
International)" Protein Seq Data Anal. 5:67-192 (1993).

The predicted mass spectra 22 are compared 24 to the
5   experimentally-derived fragment spectrum 16 to identify one or
more of the predicted mass spectra which most closely match
the experimentally-derived fragment spectrum 16. Preferably
the comparison is done automatically by calculating a
closeness-of-fit measure for each of the plurality of
10   predicted mass spectra 22 (compared to the experimentally-
derived fragment spectrum 16). It is believed that, in
general, there is high probability that the peptide analyzed
by the tandem mass spectrometer has an amino acid sequence
identical to one of the sub-sequences taken from the protein
15   sequence library 20 which resulted in a predicted mass
spectrum 22 exhibiting a high closeness-of-fit with respect to
the experimentally-derived fragment spectrum 16. Furthermore,
when the peptide analyzed by the tandem mass spectrometer 14
was derived from a protein, it is believed there is a high
20   probability that the parent protein is identical or similar to
the protein whose sequence in the protein sequence library 20
includes a sub-sequence that resulted in a predicted mass
spectra 22 which had a high closeness-of-fit with respect to
the fragment spectrum 16. Preferably, the entire procedure
25   can be performed automatically using, e.g, a computer to
calculate predicted mass spectra 22 and/or to perform
comparison 24 of the predicted mass spectra 22 with the
experimentally-derived fragment spectrum 16.

Fig. 3 is a flow diagram showing one method for
30   predicting mass spectra 22 and performing the comparison 24.
According to the method of Fig. 3, the experimentally-derived
fragment spectrum 16 is first normalized 32. According to one
normalization method, the experimentally-derived fragment
spectrum 16 is converted to a list of masses and intensities.
35   The values for the precursor ion are removed from the file.
The square root of all the intensity values is calculated and
normalized to a maximum intensity of 100. The 200 most
intense ions are divided into ten mass regions and the maximum

intensity is normalized to 100 within each region. Each ion
which is within 3.0 daltons of its neighbor on either side is
given the greater intensity value, if a neighboring intensity
is greater than its own intensity. Of course, other

5   normalizing methods can be used and it is possible to perform
analysis without performing normalization, although
normalization is, in general, preferred. For example, it is
possible to use maximum intensities with a value greater than
or less than 100. It is possible to select more or fewer than

10   the 200 most intense ions. It is possible to divide into more
or fewer than 10 mass regions. It is possible to make the
window for assuming the neighboring intensity value to be
greater than or less than 3.0 daltons.

In order to generate predicted mass spectra from a

15   protein sequence library, according to the process of Fig. 3,
sub-sequences within each protein sequence are identified
which have a mass which is within a tolerance amount of the
mass of the unknown peptide. As noted above, the mass of the
unknown peptide is known from the tandem mass spectrometer 34.

20   Identification of candidate sub-sequences 34 is shown in
greater detail in Fig. 4. In general, the process of
identifying candidate sub-sequences involves summing the
masses of linear amino acid sequences until the sum is either
within a tolerance of the mass of the unknown peptide (the

25   "target" mass) or has exceeded the target mass (plus
tolerance). If the mass of the sequence is within tolerance
of the target mass, the sequence is marked as a candidate. If
the mass of the linear sequence exceeds the mass of the
unknown peptide, then the algorithm is repeated, beginning

30   with the next amino acid position in the sequence.

According to the method of Fig. 4, a variable m,
indicating the starting amino acid along the sequence is
initialized to 0 and incremented by 1 (36, 38). The sum,
representing the cumulative mass and a variable n representing

35   the number of amino acids thus far calculated in the sum, are
initially set to 0 (40) and variable n is incremented 42. The
molecular weight of a peptide corresponding to a sub-sequence
of a protein sequence is calculated in iterative fashion by

8

steps 44 and 46. In each iteration, the sum is incremented by the molecular weight of the amino acid of the next (nth) amino acid in the sequence 44. Values of the sum 44 may be stored for use in calculating fragment masses for use in predicting a

5   fragment mass spectrum as described below. If the resulting sum is less than the target mass decremented by a tolerance 46, the value of n is incremented 42 and the process is repeated 44. A number of tolerance values can be used. In one embodiment, a tolerance value of $\pm0.05\%$ of the mass of the

10   unknown peptide was used. If the new sum is no longer less than a tolerance amount below the target mass, it is then determined if the new sum is greater than the target mass plus the tolerance amount. If the new sum is more than the tolerance amount in excess of the target mass, this particular

15   sequence is not considered a candidate sequence and the process begins again, starting from a new starting point in the sequence (by incrementing the starting point value in (38)). If, however, the sum is not greater than the target mass plus the tolerance amount, it is known that the sum is

20   within one tolerance amount of a target mass and, thus, that the sub-sequence beginning with mth amino and extending to the (m + n)th amino acid of the sequence is a candidate sequence. The candidate sequence is marked, e.g., by storing the values of m and n to define this sub-sequence.

25   Returning to Fig. 3, once a plurality of candidate sub-sequences have been identified, a fragment mass spectrum is predicted for each of the candidate sequences 52. The fragment mass spectrum is predicted by calculating the fragment ion masses for the type b- and y- ions for the amino

30   acid sequence. When a peptide is fragmented and the charge is retained on the N-terminal cleavage fragment, the resulting ion is labelled as a b-type ion. If the charge is retained on the c-type terminal fragment, it is labelled a y-type ion. Masses for type b- ions were calculated by summing the amino

35   acid masses and adding the mass of a proton. Type y- ions were calculated by summing, from the c-terminus, the masses of the amino acids and adding the mass of water and a proton to the initial amino acid. In this way, it is possible to

calculate an m/z for each fragment. However, in order to
provide a predicted mass spectrum, it is also necessary to
assign an intensity value for each fragment. It might be
possible to predict, on a theoretical basis, intensity value

5 for each fragment. However, this procedure is difficult. It
has been found useful to assign intensities in the following
fashion. The value of 50.0 is assigned to each b and y ion.
To masses of 1 dalton on either side of the fragment ion, an
intensity of 25.0 is assigned. Peak intensities of 10.0 and -

10 17.0 and -18.0 daltons below the m/z of each b- and y- ion
location (for both NH_3 and H_2O loss), and peak intensities of
10.0 and -28.0 amu of each type b ion location (for type a-
ions).

Returning to Fig. 3, after calculation of predicted
15 m/z values and assignment of intensities, it is preferred to
calculate a measure of closeness-of-fit between the predicted
mass spectra 22 and the experimentally-derived fragment
spectrum 16. A number of methods for calculating closeness-
of-fit are available. In the embodiment depicted in Fig. 3, a

20 two-step method is used 54. The two-step method includes
calculating a preliminary closeness-of-fit score, referred to
here as S_p 56 and, for the highest-scoring amino acid
sequences, calculating a correlation function 58. According
to one embodiment, S_p is calculated using the following

25 formula:

$$S_p = (\sum i_m) * n_i * (1 + \beta) * (1 - \rho) / n_t \qquad (1)$$

where i_m = matched intensities, n_i = number of matched
fragment ions, β = type b- and y-ion continuity, ρ = presence
30 of immonium ions and their respective amino acids in the
predicted sequence, n_t = total number of fragment ions. The
factor, β, evaluates the continuity of a fragment ion series.
If there was a fragment ion match for the ion immediately
preceding the current type b- or y-ion, β is incremented by
35 0.075 (from an initial value of 0.0). This increases the
preliminary score for those peptides matching a successive

series of type b- and y-ions since extended series of ions of
the same type are often observed in MS/MS spectra. The factor
p evaluates the presence of immonium ions in the low mass end
of the mass spectrum. Immonium ions are diagnostic for the

5 presence of some types of amino acids in the sequence. If
immonium ions are present at 110.0, 120.0, or 136.0 Da ($\pm$ 1.0
Da) in the processed data file of the unknown peptide with
normalized intensities greater than 40.0, indicating the
presence of histidine, phenylalanine, and tyrosine

10 respectively, then the sequence under evaluation is checked
for the presence of the amino acid indicated by the immonium
ion. The preliminary score, S_p, for the peptide is either
augmented or depreciated by a factor of $(1 - p)$ where p is the
sum of the penalties for each of the three amino acids whose

15 presence is indicated in the low mass region. Each individual
p can take on the value of -0.15 if there is a corresponding
low mass peak and the amino acid is not present in the
sequence, +0.15 if there is a corresponding low mass peak and
the amino acid is present in the sequence, or 0.0 if the low

20 mass peak is not present. The total penalty can range from
-0.45 (all three low mass peaks present in the spectrum yet
none of the three amino acids are in the sequence) to +0.45
(all three low mass peaks are present in the spectrum and all
three amino acids are in the sequence).

25 Following the calculation of the preliminary
closeness-of-fit score S_p, those candidate predicted mass
spectra having the highest S_p scores are selected for further
analysis using the correlation function 58. The number of
candidate predicted mass spectra which are selected for

30 further analysis will depend largely on the computational
resources and time available. In one embodiment, 300
candidate peptide sequences with the highest preliminary score
were selected.

For purposes of calculating the correlation

35 function, 58, the experimentally-derived fragment spectrum is
preprocessed in a fashion somewhat different from
preprocessing 32 employed before calculating S_p. For purposes
of the correlation function, the precursor ion was removed

from the spectrum and the spectrum divided into 10 sections.
Ions in each section were then normalized to 50.0. The
sectionwise normalized spectra 60 were then used for
calculating the correlation function. According to one
5   embodiment, the discrete correlation between the two functions
is calculated as:

$$R_\tau = \sum_{i=0}^{n-1} x_i y_{i+\tau} \tag{2}$$

where τ is a lag value. The discrete correlation theorem
states that the discrete correlation of two real functions x
10   and y is one member of the discrete Fourier transform pair

$$R_\tau = X_\tau Y^* \tau \tag{3}$$

where X(t) and Y(t) are the discrete Fourier transforms of
x(i) and y(i) and the Y^* denotes complex conjugation.
Therefore, the cross-correlations can be computed by Fourier
15   transformation of the two data sets using the fast Fourier
transform (FFT) algorithm, multiplication of one transform by
the complex conjugate of the other, and inverse transformation
of the resulting product. In one embodiment, all of the
predicted spectra as well as the pre-processed unknown
20   spectrum were zero-padded to 4096 data points since the MS/MS
spectra are not periodic (as intended by the correlation
theorem) and the FFT algorithm requires N to be an integer
power of two, so the resulting end effects need to be
considered. The final score attributed to each candidate
25   peptide sequence is R(0) minus the mean of the
cross-correlation function over the range -75<t<75. This
modified "correlation parameter" described in Powell and
Heiftje, Anal. Chim. Acta, Vol. 100, pp 313-327 (1978) shows
better discrimination over just the spectral correlation
30   coefficient R(0). The raw scores are normalized to 1.0. In
one embodiment, output 62 includes the normalized raw score,
the candidate peptide mass, the unnormalized correlation
coefficient, the preliminary score, the fragment ion

continuity β, the immonium ion factor ρ, the number of type b-
and y-ions matched out of the total number of fragment ions,
their matched intensities, the protein accession number, and
the candidate peptide sequence.

5   If desired, the correlation function 58 can be used
to automatically select one of the predicted mass spectra 22
as corresponding to the experimentally-derived fragment
spectrum 16. Preferably, however, a number of sequences from
the library 20 are output and final selection of a single
10   sequence is done by a skilled operator.

In addition to predicting mass spectra from protein
sequence libraries, the present invention also includes
predicting mass spectra based on nucleotide databases. The
procedure involves the same algorithmic approach of cycling
15   through the nucleotide sequence. The 3-base codons will be
converted to a protein sequence and the mass of the amino
acids summed in a fashion similar to the summing depicted in
Fig. 4. To cycle through the nucleotide sequence, a 1-base
increment will be used for each cycle. This will allow the
20   determination of an amino acid sequence for each of the three
reading frames in one pass. The scoring and reporting
procedures for the search can be the same as that described
above for the protein sequence database.

Depending on the computing and time resources
25   available, it may be advantageous to employ data-reduction
techniques. Preferably these techniques will emphasize the
most informative ions in the spectrum while not unduly
affecting search speed. One technique involves considering
only some of the fragment ions in the MS/MS spectrum. A
30   spectrum for a peptide may contain as many as 3,000 fragment
ions. According to one data reduction strategy, the ions are
ranked by intensity and some fraction of the most intense ions
(e.g., the top 200 most intense ions) will be used for
comparison. Another approach involves subdividing the
35   spectrum into, e.g., 4 or 5 regions and using the 50 most
intense ions in each region as part of the data set. Yet
another approach involves selecting ions based on the
probability of those ions being sequence ions. For example,

ions could be selected which exist in mass windows of 57
through 186 daltons (range of mass increments for the 20
common amino acids from GLY to TRP) that contain diagnostic
features of type b- or y- ions, such as losses of 17 or 18
5   daltons (NH_3 or H_2O) or a loss of 28 daltons (CO).

The techniques described above are, in general,
applicable to spectra of peptides with charged states of +1 or
+2, typically having a relatively short amino acid sequence.
Using a longer amino acid sequence increases the probability
10 of a unique match to a protein sequence. However, longer
peptide sequences have a greater likelihood of containing more
basic amino acids, and thus producing ions of higher charge
state under electro-spray ionization conditions. According to
one embodiment of the invention, algorithms are provided for
15 searching a database with MS/MS spectra of highly charged
peptides (+3, +4, +5, etc.). According to one approach, the
search program will include an input for the charge state (N)
of the precursor ion used in the MS/MS analysis. Predicted
fragment ions will be generated for each charge state less
20 than N. Thus, for peptide of +4, the charge states of +1, +2
and +3 will be generated for each fragment ion and compared to
the MS/MS spectrum.

The second strategy for use with multiply charged
spectra is the use of mathematical deconvolution to convert
25 the multiply charged fragment ions to their singly charged
masses. The deconvoluted spectrum will then contain the
fragment ions for the multiply charged fragment ions and their
singly charged counterparts.

To speed up searches of the database, a directed-
30 search approach can be used. In cases where experiments are
performed on specific organisms or specific types of proteins,
it is not necessary to search the entire database on the first
pass. Instead, a search sequence protein specific to a
species or a class of proteins can be performed first. If the
35 search does not provide reasonable answers, then the entire
database is searched.

A number of different scoring algorithms can be used
for determining preliminary closeness of fit or correlation.

In addition to scoring based on the number of matched ions multiplied by the sum of the intensity, scoring can be based on the percentage of continuous sequence coverage represented by the sequence ions in the spectrum. For example, a 10

5   residue peptide will potentially contain 9 each of b- and y-type sequence ions. If a set of ions extends from B_1 to B_9, then a score of 100 is awarded, but if a discontinuity is observed in the middle of the sequence, such as missing the B_5 ion, a penalty is assessed. The maximum score is awarded for

10   an amino acid sequence that contains a continuous ion series in both the b and y directions.

In the event the described scoring procedures do not delineate an answer, an additional technique for spectral comparison can be used in which the database is initially

15   searched with a molecular weight value and a reduced set of fragment ions. Initial filtering of the database occurs by matching sequence ions and generating a score with one of the methods described above. The resulting set of answers will then undergo a more rigorous inspection process using a

20   modified full MS/MS spectrum. For the second stage analysis, one of several spectral matching approaches developed for spectral library searching is used. This will require generating a "library spectrum" for the peptide sequence based on the sequence ions predicted for that amino acid sequence.

25   Intensity values for sequence ions of the "library spectrum" will be obtained from the experimental spectrum. If a fragment ion is predicted at m/z 256, then the intensity value for the ion in the experimental spectrum at m/z=256 will be used as the intensity of the ion in the predicted spectrum.

30   Thus, if the predicted spectrum is identical to the "unknown" spectrum, it will represent an ideal spectrum. The spectra will then be compared using a correlation function. In general, it is believed that the majority of computational time for the above procedure is spent in the iterative search

35   process. By multiplexing the analysis of multiple MS/MS spectra in one pass through the database, an overall improvement in efficiency will be realized. In addition, the mass tolerance used in the initial pre-filtering can affect

search times by increasing or decreasing the number of
sequences to analyze in subsequent steps. Another approach to
speed up searches involves a binary encryption scheme where
the mass spectrum is encoded as peak/no peak at every mass
5   depending on whether the peak is above a certain threshold
value. If intensive use of a protein sequence library is
contemplated, it may be possible to calculate and store
predicted mass values of all sub-sequences within a
predetermined range of masses so that at least some of the
10   analysis can be performed by table look-up rather than
calculation.

          Figs. 6A-6E are flow charts showing an analysis
procedure according to one embodiment of the present
invention. After data is acquired from the tandem mass
15   spectrometer, as described above 602, the data is saved to a
file and converted to an ASCII format 604. At this point, a
preprocessing procedure is started 606. The user enters
information regarding the peptide mass in the precursor ion
charge state 608. Mass/intensity values are loaded from the
20   ASCII file, with the values being rounded to unit masses 610.
The previously-identified precursor ion contribution of this
data is removed 612. The remaining data is normalized to a
maximum intensity of 100 614. At this point, different paths
can be taken. In one case, the presence of any immonium ions
25   (H, F and Y) is noted 616 and the peptide mass and immonium
ion information is stored in a datafile 618. In another
route, the 200 most intense peaks are selected 620. If two
peaks are within a predetermined distance (e.g., 2 amu) of
each other, the lower intensity peak is set equal to a greater
30   intensity 622. After this procedure, the data is stored in a
datafile for preliminary scoring 624. In another route, the
data is divided into a number of windows, for example ten
windows 626. Normalization is performed within each window,
for example, normalizing to a maximum intensity of 50 628.
35   This data is then stored in a datafile for final correlation
scoring 630. This ends the preprocessing phase, according to
this embodiment 632.

The database search is started 634 and the search
parameters and the data obtained from the preprocessing
procedure (Fig. 6A) are loaded 636. A first batch of database
sequences is loaded 638 and a search procedure is run on a
5   particular protein 640. The search procedure is detailed in
Fig. 6C. As long as the end of the batch has not been reached
the index is incremented 642 and the search routine is
repeated 640. Once it is determined that the end of a batch
has been reached 644, as long as the end of the database has
10   not been reached, the second index 646 is incremented and a
new batch of database sequences is loaded 638. Once the end
of the database has been reached 628, a correlation analysis
is performed 630 (as detailed in Fig. 6E), the results are
printed 632 and the procedure ends 634.

15   When the search procedure is started 638 (Fig. 6C),
an index I1 is set to zero 646 to indicate the start position
of the candidate peptide within the amino acid being searched
640. A second index I2, indicating the end position of the
candidate peptide within the amino acid being searched, is
20   initially set equal to I1 and the variable Pmass, indicating
the accumulated mass of the candidate peptide is initialized
to zero 648. During each iteration through a given candidate
peptide 650 the mass of the amino acid at position I2 is added
to Pmass 652. It is next determined whether the mass thus-far
25   accumulated (Pmass) equals the input mass (i.e., the mass of
the peptide) 654. In some embodiments, this test may be
performed as plus or minus a tolerance rather than requiring
strict equality, as noted above. If there is equality
(optionally within a tolerance) an analysis routine is started
30   656 (detailed in Fig. 6D). Otherwise, it is determined
whether Pmass is less than the input mass (optionally within a
tolerance). If not, the index I2 is incremented 658 and the
mass of the amino acid at the next position (the incremented
I2 position) is added to Pmass 652. If Pmass is greater than
35   the input mass (optionally by more than a tolerance 660) it is
determined whether index I1 is at the end of a protein 662.
If so, the search routine exits 664. Otherwise, index I1 is
incremented 666 so that the routine can start with a new start

position of a candidate peptide and the search procedure returns to block 648.

When the analysis procedure is started 670 (Fig. 6D), data indicative of b- and y- ions for the candidate
5   peptide are generated 672, as described above. It is determined whether the peak is within the top 200 ions 674. The peak intensity is summed and the fragmented match index is incremented 676. If previous b- or y- ions are matched 678, the β index is incremented 680. Otherwise, it is determined
10   whether all fragment ions have been analyzed. If not, the fragment index is incremented 684 and the procedure returns to block 674. Otherwise, a preliminary score such as S_p, described above is calculated 686. If the newly-calculated S_p is greater than the lowest score 688 the peptide sequence is
15   stored 690 unless the sequence has already been stored, in which case the procedure exits 692.

At the beginning of the correlation analysis (Fig. 6E), a stored candidate peptide is selected 693. A theoretical spectrum for the candidate peptide is created 694,
20   correlated with experimental data 695 and a final correlation score is obtained 696, as described above. The index is incremented 697 and the process repeated from block 693 unless all candidate peptides have been scored 698, in which case the correlation analysis procedure exits 699.

25   The following examples are offered by way of illustration, not limitation.

Experimental
Example #1

30   MHC complexes were isolated from HS-EBV cells transformed with HLA-DRB*0401 using antibody affinity chromatography. Bound peptides were released and isolated by filtration through a Centricon 10 spin column. The heavy chain of glycosaparginase from human leukocytes was isolated.
35   Proteolytic digestions were performed by dissolving the protein in 50 mM ammonium bicarbonate containing 10 mM Ca^{++}, pH 8.6. Trypsin was added in a ratio of 100/1 protein/enzyme.

Analysis of the resulting peptide mixtures was
performed by LC-MS and LC-MS/MS. Briefly, molecular weights
of peptides were recorded by scanning Q3 or Q1 at a rate of
400 Da/sec over a mass range of 300 to 1600 throughout the
5   HPLC gradient. Sequence analysis of peptides was performed
during a second HPLC analysis by selecting the precursor ion
with a 6 amu (FWHH) wide window in Q_1 and passing the ions
into a collision cell filled with argon to a pressure of 3-5
mtorr. Collision energies were on the order of 20 to 50 eV.
10   The fragment ions produced in Q_2 were transmitted to Q_3 and a
mass range of 50 Da to the molecular weight of the precursor
ion was scanned at 500 Da/sec to record the fragment ions.
The low energy spectra of 36 peptides were recorded and stored
on disk. The genpept database contains protein sequences
15   translated from nucleotide sequences. A text search of the
database was performed to determine if the sequences for the
peptide amino acid sequences used in the analysis were present
in the database. Subsequently, a second database was created
from the whole database by appending amino acid sequences for
20   peptides not included.

        The spectrum data was converted to a list of masses
and intensities and the values for the precursor ion were
removed from the file. The square root of all the intensity
values was calculated and normalized to a maximum intensity of
25   100.0. All ions except the 200 most intense ions were removed
from the file. The remaining ions were divided into 10 mass
regions and the maximum intensity normalized to 100.0 within
each region. Each ion within 3.0 daltons of its neighbor on
either side was given the greater intensity value, if the
30   neighboring intensity was greater than its own intensity.
This processed data was stored for comparison to the candidate
sequences chosen from the database search. The MS/MS spectrum
was modified in a different manner for calculation of a
correlation function. The precursor ion was removed from the
35   spectrum and the spectrum divided into 10 equal sections.
Ions in each section were then normalized to 50.0. This
spectrum was used to calculate the correlation coefficient

against a predicted MS/MS spectrum for each amino acid
sequence retrieved from the database.

Amino acid sequences from each protein were
generated by summing the masses, using average masses for the
5   amino acids, of the linear amino acid sequence from the amino
terminus (n). If the mass of the linear sequence exceeded the
mass of the unknown peptide, then the algorithm returned to
the amino terminal amino acid and began summing amino acid
masses from the n+1 position. This process was repeated until
10 every linear amino acid sequence combination had been
evaluated. When the mass of the amino acid sequence was
within ±0.05% (minimum of ±1 Da) of the mass of the unknown
peptide, the predicted m/z values for the type b- and y-ions
were generated and compared to the fragment ions of the
15 unknown sequence. A preliminary score (S_p) was generated and
the top 300 candidate peptide sequences with the highest
preliminary score were ranked and stored. A final analysis of
the top 300 candidate amino acid sequences was performed with
a correlation function. Using this function a theoretical
20 MS/MS spectrum for the candidate sequence was compared to the
modified experimental MS/MS spectrum. Correlation
coefficients were calculated, ranked and reported. The final
results were ranked on the basis of the normalized correlation
coefficient.
25   The spectrum shown in Fig. 5 was obtained by
LC-MS/MS analysis of a peptide bound to a DRB*0401 MHC class
II complex. A search of the genpept database containing
74,938 protein sequences identified 384,398 peptides within a
mass tolerance of ±0.05% (minimum of ±1Da) of the molecular
30 weight of this peptide. By comparing fragment ion patterns
predicted for each of these amino acid sequences to the
pre-processed MS/MS spectra and calculating a preliminary
score, the number of candidate sequences was cutoff at 300. A
correlation analysis was then performed with the predicted
35 MS/MS spectra for each of these sequences and the modified
experimental MS/MS spectrum. The results of the search
through the genpept database with the spectrum in Fig. 5 are
displayed in Table 1. Two peptides of similar sequence,

DLRSWTAADAAQISK, DLRSWTAADAAQISQ, were identified as the
highest scoring sequences (C_n values). Their correlation
coefficients are identical so their rankings in Table 1 are
arbitrary. The amino acid sequence DLRSWTAADAAQISK occurs in

5   five proteins in the genpept database while the sequence
DLRSWTAADAAQISQ occurs in only one. The top three sequences
appear in immunologically related proteins and the rest of the
proteins appear to have no correlation to one another. A
second search using the same MS/MS spectrum was performed with

10   the *Homo sapiens* subset of the genpept database to compare the
results. These data are presented in Table 2. In both
searches the correct sequence tied for the top position. Both
amino acid sequences have identical correlation coefficients,
C_n, although the sequences differ by Lys and Gln at the

15   C-terminus. These two amino acids have the same nominal mass
and would be expected to produce similar MS/MS spectra. The
sum of the normalized fragment ion intensities, I_m, for the
matched fragment ions for the two peptides are different with
the correct sequence having the greater value. The correct

20   sequence also matched an additional fragment ion in the
preliminary scoring procedure identifying 70% of the predicted
fragment ions for this amino acid sequence in the
pre-processed spectrum. These matches are determined as part
of the preliminary scoring procedure.

25

Table 1

NO	Mass	C_n	C	S_p	β	ρ	Ions	I_m	Accession No	No of Proteins	Sequence
1	1734.90	1.000	4.203	1809.9	1.125	0.000	20/30	1277.61	X17273	+5	DLRSWTAADTAAQISK
2	1734.86	1.000	4.203	2020.1	1.200	0.000	21/30	1311.79	M84380	+5	DLRSWTAADTAAQISQ
3	1733.91	0.893	3.752	1440.8	0.975	0.000	19/30	1151.87	X60251	+2	DLRSWTAADTAAQVTQ
4	1733.96	0.727	3.057	606.1	0.525	0.000	14/30	851.71	m90516		VIFLBDDVAAVVDGR
5	1733.96	0.685	2.881	1012.5	0.750	0.000	16/28	1012.47	D13262	+1	TDMBLLAGIWQDVL
6	1734.94	0.684	2.874	805.8	0.675	0.000	16/32	962.12	S58174		RCLTASTLGLTQRVNVN
7	1743.96	0.683	2.869	763.4	0.600	0.000	15/30	954.28	X17149		GSSLMABDNLBLAVRK
8	1733.97	0.767	2.839	790.1	0.675	0.000	17/36	998.90	M19085		SGILTLSGGLSSNABVTAG
9	1733.93	0.675	2.836	682.9	0.525	0.000	13/26	895.55	M29146	+1	YPSKQINBLMBNVL
10	1734.90	0.668	2.809	590.8	0.525	0.000	15/36	929.73	M84615		AKSGBANVTSATOTIGOTI
11	1733.99	0.663	2.786	889.5	0.750	0.000	16/32	1016.53	D13297	+2	GHENNKSAASSKGKITL
12	1734.94	0.648	2.724	459.8	0.450	0.000	13/30	731.82	M32066		MLLDAIKGGSSDLMFE
13	1735.89	0.639	2.685	1461.3	0.900	0.000	19/30	1214.40	M28242	+1	PTPBSVSRLLEKISAG
14	1733.95	0.636	2.672	614.5	0.600	0.000	13/28	827.23	J05110		SKPKTMNVLARLDV
15	1733.99	0.632	2.658	501.7	0.525	0.000	12/30	865.03	M24378		MTWTDSMAGVNNSGSQ
16	1733.93	0.628	2.641	563.5	0.300	0.000	14/30	928.90	M76547		KTDSMGHIKLDCPSLK
17	1734.91	0.626	2.629	580.6	0.600	0.000	13/28	781.63	M30149	+5	BMLLMBRGYGKLVD
18	1734.90	0.622	2.613	756.6	0.525	0.000	17/34	992.25	M33962		SGSL8PBHGPIVVHCSAG
19	1734.90	0.619	2.600	625.6	0.450	0.000	14/30	924.55	K02670		LPVLLAVPLAGLMIGR
20	1734.91	0.617	2.593	563.2	0.600	0.000	12/26	762.63	M63585	+1	DILDBLPMNQDGR
21	1735.86	0.617	2.592	491.3	0.525	0.000	13/30	743.39	X67813		AQLISAYSLVDPEKAK
22	1735.05	0.615	2.585	472.5	0.600	0.000	12/28	689.08	J04120	+2	QBLVINIGWIIBGMP
23	1734.94	0.615	2.583	548.0	0.450	0.000	13/30	872.19	X54240		VDBSIWEDMSVVSLSQ
24	1734.63	0.609	2.561	493.4	0.525	0.000	12/28	754.92	X17637		DVLANAYRISRQBAQ
25	1734.99	0.603	2.536	602.6	0.450	0.000	16/23	831.23	D90034		CBDSLDERKIKGVIB

22

Table 1 (continued)

Mass = calculated candidate peptide mass, C_n = normalized correlation coefficient, C = correlation coefficient, S_p = preliminary score, β = fragment ion continuity, ρ = immonium ions, Ions = number of type b- and y-ions matched out of the total number of fragment ions for candidate sequence, Im = matched fragment intensities, Accesion No = genpept accession number, and the candidate peptide sequence, mass = 1734.90, frag tolerance = 0.75, mass tolerance = 1.000.

Table 2

No	Mass	C_n	C	S_p	β	ρ	Ions	I_m	Accession No	No of Proteins	Sequence
1	1734.90	1.000	4.203	1809.9	1.125	0.000	20/30	1277.61	X17273	+5	DLRSWTAADTAAQISK
2	1734.86	1.000	4.203	2020.1	1.200	0.000	21/30	1311.79	M84380		DLRSWTAADTAAQISQ
3	1733.96	0.727	3.057	606.1	0.525	0.000	14/30	851.71	M90516		VIFLBDDDVAAVVDGR
4	1733.96	0.598	2.514	642.1	0.600	0.000	15/32	856.16	X54637		GSSGRNPQASLPGKKAK
5	1733.99	0.577	2.424	465.9	0.275	0.000	12/28	790.70	X02598		CKLQSGTHICLMTDQL
6	1733.93	0.558	2.347	674.2	0.450	0.000	15/32	991.85	Z18330	+5	MTQSPATLSVSPGBRAT
7	1734.12	0.554	2.329	300.2	0.225	0.000	10/28	686.18	Z22658		PTVGKBLNNIPKLPD
8	1734.96	0.553	2.323	447.6	0.375	0.000	13/32	801.32	X02763		SSGILSRSSVGPCIRSQ
9	1734.94	0.549	2.307	375.1	0.225	0.000	12/30	765.57	J03259		PVVDYTBLVLTGNNLT
10	1735.89	0.545	2.292	544.9	0.375	0.000	14/30	849.24	M18728	+1	PASPAWQDDAVISISQ
11	1735.86	0.543	2.281	590.0	0.450	0.000	13/28	876.39	J03202	+2	BTLENBANNIRGEAB
12	1734.91	0.542	2.276	646.1	0.450	0.000	15/30	891.21	X51592		KCBIQAALBBABASLB
13	1734.83	0.537	2.259	574.0	0.450	0.000	13/26	791.73	X63575	+1	MTNSDFYSNQRNB
14	1733.93	0.534	2.243	500.9	0.525	0.000	13/30	758.05	X15875		LTQWADQSTBPALSQI
15	1733.97	0.532	2.235	514.3	0.600	0.000	13/30	741.72	S79307		LRLSSVTAADTAVYYC
16	1735.86	0.526	2.210	377.1	0.300	0.000	12/28	676.80	X12530	+2	SRKNSPSTOYCYSIQ
17	1733.96	0.525	2.206	518.3	0.525	0.000	15/34	770.42	M25668		TAAGGBSALAPSVFKQAK
18	1734.99	0.525	2.205	360.3	0.225	0.000	12/30	735.30	M81830		SSGTRVGSSKRKGBK
19	1733.99	0.524	2.202	376.7	0.300	0.000	12/30	724.46	K03456	+2	SBGATPQDLNNKLNIV
20	1733.94	0.522	2.195	863.6	0.675	0.000	15/28	962.41	M92269	+1	ECLKRQKDRGGDISQ
21	1734.94	0.520	2.187	603.4	0.525	0.000	15/36	949.64	M85289		AKGSVYIGGAPDVATLTQ
22	1734.85	0.520	2.187	540.6	0.450	0.000	13/28	802.95	M17081	+4	KTDIKVVDRDSEBAE
23	1734.99	0.516	2.167	526.1	0.750	0.000	13/28	647.52	L07217	+5	IIRSENFTNNAKTII
24	1733.99	0.515	2.166	437.6	0.450	0.000	12/30	754.47	X70848	+3	VDBKIPCCHGGLSIVDL
25	1735.67	0.510	2.143	532.9	0.450	0.000	13/30	848.14	M95929		SGSDTPQQDNDQLNSB

Example 12

To examine the complexity of the mixture of peptides
obtained by proteolysis of the total proteins from *S.
cerevisiae* cells, 10^8 cells were grown and harvested. After
5   lysis, the total proteins were contained in ~9 mL of solution.
A 0.5 mL aliquot was removed for proteolysis with the enzyme
trypsin. From this solution two microliters were directly
injected onto a micro-LC (liquid chromatography) column for MS
analysis. In a complex mixture of peptides it is conceivable
10   that multiple peptide ions may exist at the same m/z and
contribute to increased background, complicating MS/MS
analysis and interpretation. To test the ability to obtain
sequence information by MS/MS from these complex mixtures of
peptides, ions from the mixture were selected with on-line
15   MS/MS analysis. In no case were the spectra contaminated with
fragment ions from other peptides. A partial list of the
identified sequences is presented in Table 3.

Table 3

S. cerevisiae Protein	Amino acid Sequence
enolase	DPPAEDDWEAWSH
hypusine containing protein HP2	APEGELGDSLQTAFDEGK
phosphoglycerate kinase	TGGGASLELLEGK
BMH1 gene product	QAPDDAIAELDTLSEESYK
pyruvate kinase	IPAGWQGLDNGPSER
phosphoglycerate kinase	LPGTDVDLPALSEK
hexokinase	IEDDPPENLEDTDDDPQK
enolase	EEALDLIVDAIK
enolase	NPTVEVELTTEK

The MS/MS spectra presented in Table 1 were
interpreted using the described database searching method.
40   This method serves as a data pre-filter to match MS/MS spectra
to previously determined amino acid sequences. Pre-filtering
the data allows interpretation efforts to be focused on
previously unknown amino acid sequences. Results for some of
the MS/MS spectra are shown in Table 4. No pre-assigning of
45   sequence ions or manual interpretation is required prior to
the search. However, the sequences must exist in the

database. The algorithm first pre-processed the MS/MS data
and then compared all the amino acid sequences in the database
within ±1 amu of the mass of the precursor ion of the MS/MS
spectrum. The predicted fragmentation patterns of the amino

5 acid sequences within the mass tolerance were compared to the
experimental spectrum. Once an amino acid sequence was within
this mass tolerance, a final closeness-of-fit measure was
obtained by reconstructing the MS/MS spectra and performing a
correlation analysis to the experimental spectrum. Table 4

10 lists a number of spectra used to test the efficacy of the
algorithm.

The computer program described above has been modified to
analyze the MS/MS spectra of phosphorylated peptides. In this
algorithm all types of phosphorylation are considered such as

15 Thr, Ser, and Tyr. Amino acid sequences are scanned in the
database to find linear stretches of sequence that are
multiples of 80 amu below the mass of the peptide under
analysis. In the analysis each putative site of
phosphorylation is considered and attempts to fit a

20 reconstructed MS/MS spectrum to the experimental spectrum are
made.

Table 4

List of results obtained searching genpept and species specific databases using MS/MS spectra for the respective peptides.

No.	Mass	Amino Acid Sequence of Peptides used in the Search	Genpept Database	Genpept Database[3]	Species Specific
1	1734.9	DLRSWTAADTAAQISQ	1	1	1
2	1749	DLRSWTAADTAAQITQ	1	1	1
3	1186.5	MATPLLMQALP	-	-	13
4	1317.7	MATPLLMQALP	61	61	17
5	1571.6	EGVNDNEEGFFSAR[1,2]	1*	1	1
6	1571.6	EGVNDNEEGFFSAR[1,2]	1*	1	1
7	1297.5	DRVYIHPFHL(+2)	1	1	1
8	1297.5	DRVYIHPFHL(+2)	2	2	2
9	1297.5	DRVYIHPFHL(+3)	1	1	1
10	1593.8	VEADVAGHGQDILIR[2]	1	1	1
11	1393.7	HGVTVLTALGAILK[2]	1	1	1
12	1741.8	HSGQAEGYSYTDANIK[2]	1	1	1
13	848.8	HSGQAEGY[2](+1)	1	1	1
14	723.9	MAPGGLK[2,3](+1)	-	-	-
15	636.8	GATLPK[2](+1)[QATLPG, KTLPK]	1*	1*	6
16	524.6	TEFK(+1)	1*	1*	5
17	1251.4	DRNDLLTYLK[1,2]	5*	5	1
18	1194.4	VLVLDTDYKK[2]	6*	6	2
19	700.7	CRGDSY[1](CGRDSY)	3*	1	1
20	700.7	CRGDSY[1](+1)	-	-	7
21	764.9	KGATLPK[2]	3	3	1
22	1169.3	TGPNLHGLFGR	1	1	1
23	1047.2	DRVYIHPF	-	-	7
24	1139.3	TLLVGESATTF(+1)	1	1	1
25	1189.4	RNVIPDSKY	1	1	1
26	613.7	SSPLPL(+1)	2	4	2
27	1323.5	LARNCQPNYW(C-161.17)	1	1	1
28	2496.7	AQSMGFINEDLSTSAQALMSDW	1	1	1
29	1551.8	VTLIHPIAMDDGLR	3	3	1
30	1803.0	GGDTVTLNETDLTQIPK	2	2	1
31	1172.4	VGEEVEIVGIK	1	1	1
32	2148.5	GWQVPAFTLGGEATDIVVMR	1	1	1
33	2553.9	VASISLPTSCASAGTQCLISGWGNTK[1]	-	1	1
34	1154.3	SSGTSYPDVLK[1]	-	3	1
35	1174.5	TLNNDIMLIK	1	1	1
36	2274.6	SIVHPSYNSNTLNNDIMLIK[1]	-	2	1

[1] not present in the genpept database
[2] sequence appended to the human database, not originally in human database
[3] amino acid sequences added to database
(-) not in the top 100 answers
* peptide of similar sequence identified

Example #1

Much of the information generated by the genome projects will be in the form of nucleotide sequences. Those stretches of nucleotide sequence that can be correlated to a gene will be translated to a protein sequence and stored in a specific database (genpept). The un-translated nucleotide

sequences represent a wealth of data that may be relevant to protein sequences. The present invention will allow searching the nucleotide database in the same manner as the protein sequence databases. The procedure will involve the same

5 algorithmic approach of cycling through the nucleotide sequence. The three-base codon will be converted to a protein sequence and the mass of the amino acids summed. To cycle through the nucleotide sequence, a one-base increment will be used for each cycle. This will allow the determination of an

10 amino acid sequence for each of the three reading frames in one pass. For example, an MS/MS spectrum is generated for the sequence Asp-Leu-Arg-Ser-Trp-Thr-Ala ((M+H)+=848) the algorithm will search the nucleotide sequence in the following manner.

15

Nucleotide sequence from the database.

nucleotides	GCG	AUC	UCC	GGU	CUU	GGA	CUG	CUC		
First pass through the sequence.										
nucleotides	GCG	AUC	UCC	GGU	CUU	GGA	CUG	CUC		Mass
amino acids	Ala	Ile	Ser	Gly	Leu	Gly	Leu	Leu		743
Second pass through the sequence.										
nucleotides	G	CGA	UCU	CCG	GUC	UUG	GAC	UGC	UC	Mass
amino acids		Arg	Ser	Pro	Val	Leu	Gly	Leu		741
Third pass through the sequence.										
nucleotides	GC	GAU	CUC	CGG	UCU	UGG	ACU	GCU	C	Mass
amino acids		Asp	Leu	Arg	Ser	Trp	Thr	Ala		848
Fourth pass through the sequence.										
nucleotides	GCG	AUC	UCC	GGU	CUU	GGA	CUG	CUC		Mass
amino acids		Ile	Ser	Gly	Leu	Gly	Leu	Leu		672

As the sequence of amino acids match the mass of the peptide the predicted sequence ions will be compared to the MS/MS spectrum. From this point on the scoring and reporting procedures for the search will be the same as for a protein

35 sequence database.

In light of the above description, a number of advantages of the present invention can be seen. The present invention permits correlating mass spectra of a protein, peptide or oligonucleotide with a nucleotide or protein

40 sequence database in a fashion which is relatively accurate, rapid, and which is amenable to automation (i.e., to operation

28

without the need for the exercise of human judgment). The
present invention can be used to analyze peptides which are
derived from a mixture of proteins and thus is not limited to
analysis of intact homogeneous proteins such as those
5   generated by specific and known proteolytic cleavage.

A number of variations and modifications of this
invention can also be used. The invention can be used in
connection with a number of different proteins or peptide
sources and it is believed applicable to any analysis using
10 mass spectrometry and proteins. In addition to the examples
described above, the present invention can be used for, for
example, monitoring fermentation processes by collecting
cells, lysing the cells to obtain the proteins, digesting the
proteins, e.g. in an enzyme reactor, and analyzing by Mass
15 spectrometry as noted above. In this example, the data could
be interpreted using a search of the organism's database
(e.g., a yeast database). As another example, the invention
could be used to determine the species of organism from which
a protein is obtained. The analysis would use a set of
20 peptides derived from digestion of the total proteins. Thus,
the cells from the organism would be lysed, the proteins
collected and digested. Mass spectrometry data would be
collected with the most abundant peptides. A collection of
spectra (e.g., 5 to 10 spectra) would be used to search the
25 entire database. The spectra should match known proteins of
the species. Since this method would use the most abundant
proteins in the cell, it is believed that there is a high
likelihood the sequences for these organisms would be
sequenced and in the database. In one embodiment, relatively
30 few cells could be used for the analysis (e.g., on the order
of $10^4 - 10^5$).

The present invention can be used in connection with
diagnostic applications, such as described for Example No. 2
above. Another example would involve identifying virally
35 infected cells. Success of such an approach is believed to
depend on the relative abundance of the viral proteins versus
the cellular proteins, at least using present equipment. If
an antibody were produced to a specific region of a protein

common to certain pathogens, the mixture of proteins could be
partially fractionated by passing the material over an
immunoaffinity column. Bound proteins are eluted and
digested. Mass spectrometry generates the data to search a

5   database. One important element is finding a general handle
to pull proteins from the cell. This approach could also be
used to analyze specific diagnostic proteins. For example, if
a certain protein variant is known to be present in some form
of cancer or genetic disease, an antibody could be produced to

10   a region of the protein that does not change. An
immunoaffinity column could be constructed with the antibody
to isolate the protein away from all the other cellular
proteins. The protein would be digested and analyzed by
tandem mass spectrometry. The database of all the possible

15   mutations in the protein could be maintained and the
experimental data analyzed against this database.

One possible example would be cystic fibrosis. This
disease is characterized by multiple mutations in CFTR
protein. One mutation is responsible for about 70% of the

20   cases and the other 30% of the cases result from a wide
variety of mutations. To analyze these mutations by genetic
testing would require many different analyses and probes. In
the assay described above, the protein would be isolated and
analyzed by tandem mass spectrometry. All the mutations in

25   the protein could be identified in an assay based on
structural information. The database used would preferably
describe all the known mutations. Implementation of this
approach is believed to involve significant difficulties. The
amount of protein required could be so large that it would be

30   impractical to obtain from a patient. This problem may be
overcome as the sensitivity of mass spectrometry improves in
the future. A protein such as CFTR is a transmembrane
protein, which are typically very difficult to manipulate and
digest. The approach described could be used for any

35   diagnostic protein. The data would be highly specific and the
data analysis essentially automated.

It is believed that the present invention can be
used with any size peptide. The process requires that

peptides be fragmented and there are methods for achieving
fragmentation of very large proteins. Some such techniques
are described in Richard D. Smith et, al., "Collisional
Activation and Collision-Activated Dissociation of Large

5   Multiply Charged Polypeptides and Proteins Produced by
Electrospray Ionization" *J. American Society for Mass
Spectrometry* (1990) Vol. I, pp. 53-65. It is believed the
present method could be used to analyze data derived from
intact proteins. Although, as noted above, it is believed

10 that there is no theoretical or absolute practical limit to
the size of peptides that could be analyzed according to this
invention, analysis using the present invention has been
performed on peptides at least in the size range from about
400 amu (4 residues) to about 2500 amu (26 residues).

15   Although in one described embodiment, candidate sub-
sequences are identified and fragment spectra are predicted as
they are needed, at the time of doing the analysis. It would
be possible, if sufficient computational resources and storage
facilities are available to perform some or all of the

20 calculations needed for candidate sequence identification
(such as calculation of sub-sequence masses) and/or spectra
prediction (such as calculation of fragment masses) and
storage of these items in a database so that some or all of
these items can be looked up rather than calculated each time

25 they are needed.

        While the present invention has been described by
way of the preferred embodiment and certain variations and
modifications, other variations and modifications of the
present invention can also be used, the invention being

30 described by the following claims.

WHAT IS CLAIMED IS:

1. A method for correlating a peptide fragment mass spectrum with amino acid sequences derived from a database of sequences, comprising:

storing data representing a first mass spectrum of a plurality of fragments of at least a first peptide;

calculating a plurality of predicted mass spectra of at least a portion of a plurality of said sequences in said database of sequences; and

calculating at least a first measure for each of said plurality of predicted mass spectra, said first measure being an indication of the closeness-of-fit between said first mass spectrum and each of said plurality of mass spectra.

2. A method, as claimed in claim 1, wherein said first mass spectrum is provided from a tandem mass spectrometer device.

3. A method, as claimed in claim 2, wherein the tandem mass spectrometer is one of a triple quadrupole mass spectrometer, a Fourier-transform cyclotron resonance mass spectrometer, a tandem time-of-flight mass spectrometer and a quadrupole ion trap mass spectrometer.

4. A method, as claimed in claim 1, wherein said database of sequences is a database of amino acid sequences of a plurality of proteins.

5. A method, as claimed in claim 1, wherein said database of sequences is a nucleotide database.

6. A method, as claimed in claim 1, further comprising selecting a first plurality of sub-sequences from said database of sequences, wherein said step of calculating a plurality of predicted mass spectra includes calculating at least one predicted mass spectrum for each of said selected first plurality of sub-sequences.

32

7. A method, as claimed in claim 1, wherein said step of calculating a first measure includes selecting those values from said first mass spectrum having an intensity greater than a predetermined threshold.

5

8. A method, as claimed in claim 1, further comprising normalizing said first spectrum prior to said step of calculating at least a first measure.

10

9. A method, as claimed in claim 1, wherein said step of calculating a plurality of predicted mass spectra includes calculating predicted mass spectra for only a portion of said sequence database.

15

10. A method, as claimed in claim 9, wherein said first peptide is derived from a protein which is obtained from a first organism and wherein said protein of said sequence database is the portion containing sequences for proteins found in said first organism.

20

11. A method, as claimed in claim 2 wherein a first mass spectrometer of said tandem mass spectrometer device is used to separate-out a component having a first mass, an activation device of said mass spectrometer device is used to

25 fragment said first component and a second mass spectrometer of said tandem mass spectrometer device is used provide said first mass spectrum.

12. A method, as claimed in claim 1, wherein said
30 first peptide is isolated by chromatography.

13. A method, as claimed in claim 1, wherein said data representing said first mass spectrum includes a plurality of mass-charge pairs.

35

14. A method, as claimed in claim 1, wherein said step of calculating predicted mass spectra comprises:

deriving a plurality of masses from portions of said plurality of sequences, each mass equal to the mass of a peptide fragment which corresponds to a portion of a sequence in said plurality of sequences;

5   selecting those masses, among said plurality of masses, which are within a predetermined mass tolerance of the mass of said first peptide and storing an indication of which portion of which sequence each of said selected masses corresponds to, to provide a plurality of candidate sequence

10   portions; and

calculating a plurality of mass-charge pairs for each of said candidate sequence portions, each of said mass-charge pairs having a mass substantially equal to the mass of a peptide fragment corresponding to a portion of one of said

15   candidate sequence portions.

15. A method, as claimed in claim 1, wherein said first measure comprises a correlation coef cient.

20   16. A method, as claimed in claim ?, wherein said step of calculating a first measure comprist.

calculating a preliminary score for each of said plurality of candidate sequence portions;

identifying a plurality of primary candidate

25   portions which have a preliminary score which is greater than at least one candidate sequence which is not identified as a primary candidate portion; and

calculating a correlation coefficient for each of said primary candidate portions.

30

17. A method, as claimed in claim 8, wherein each of said plurality of mass spectra and said first mass spectrum includes a plurality of mass-charge pairs, each mass-charge pair having an intensity value, and further comprising the

35   step of identifying, for each of said plurality of mass spectra, a set of matched fragments which have less than a predetermined difference from corresponding fragments in said first mass spectrum; and

wherein said preliminary score is the number of
fragments of a predicted mass spectrum in said set of matched
fragments multiplied by the sum of the intensity values for
the mass-charge pairs corresponding to said matched fragments.

5

18. A method for interpreting the mass spectrum of
an oligonucleotide comprising:
 providing a library of nucleotide sequences;
 storing, in a database, a plurality of nucleotide
10   sub-sequences from said library, said plurality including all
sequences smaller than n-mers;
 storing data representing a first mass spectrum of a
plurality of fragments of said oligonucleotide;
 calculating predicted mass spectra for each of said
15   plurality of nucleotide sub-sequences; and
 calculating at least a first closeness-of-fit
measure for each of said predicted mass spectra, with respect
to said first mass spectrum.

20   19. A method, as claimed in claim 14, wherein n is
10.

20. A method for determining whether a peptide in a
mixture of proteins is homologous to a portion of any of a
25   plurality of proteins specified by an amino acid sequence in a
database of sequences, comprising:
 using a tandem mass spectrometer to receive a
plurality of peptides obtained from said mixture of proteins,
to select at least a first peptide from said mixture of
30   peptides, to fragment said first peptide and to generate a
peptide fragment mass spectrum;
 storing data representing said first mass spectrum;
and
 correlating said mass spectrum with an amino acid
35   sequence in said database of sequences, to determine the
correspondence of a protein specified in said sequence
database with a protein in said mixture of proteins.

21. A method, as claimed in claim 20, wherein said step of correlating includes predicting at least one mass spectrum from said amino acid sequence.

USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO IDENTIFY AMINO ACID SEQUENCES IN DATABASES

5

ABSTRACT OF THE DISCLOSURE

A method for correlating a peptide fragment mass spectrum with amino acid sequences derived from a database is provided. A peptide is analyzed by a tandem mass spectrometer

10 to yield a peptide fragment mass spectrum. A protein sequence database or a nucleotide sequence database is used to predict one or more fragment spectra for comparison with the experimentally-derived fragment spectrum. In one embodiment, sub-sequences of the sequences found on the database which

15 define a peptide having a mass substantially equal to the mass of the peptide analyzed by the tandem mass spectrometer are identified as candidate sequences. For each candidate sequence, a plurality of fragments of the sequence are identified and the masses and m/z ratios of the fragments are

20 predicted and used to form a predicted mass spectrum. The various predicted mass spectra are compared to the experimentally derived fragment spectrum using a closeness-of-fit measure, preferably calculated with a two-step process, including a calculation of a preliminary score and, for the

25 highest-scoring predicted spectra, calculation of a correlation function.

F:\PERFSOL\DOCS\P\28877.1

DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I declare that:

My residence, post office address and citizenship are as stated below next to my name: I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural inventors are named below) of the subject matter which is claimed and for which a patent is sought on the invention entitled: **USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO IDENTIFY AMINO ACID SEQUENCES IN PROTEIN DATABASES** the specification of which __X__ is attached hereto or _____ was filed on _____ as Application Serial No. _____ and was amended on _____. (if applicable).

I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above. I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, Section 1.56. I claim foreign priority benefits under Title 35, United States Code, Section 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)

Country	Application No.	Date of Filing	Priority Claimed Under 35 USC 119	
			Yes ___	No ___
			Yes ___	No ___

I claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, Section 1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

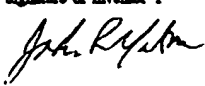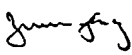| Application Serial No.	Date of Filing	Status		
		___ Patented	___ Pending	___ Abandoned
		___ Patented	___ Pending	___ Abandoned

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith.

Richard L. Hughes, Reg. No. 31,264
Steven W. Parmelee, Reg. No. 31,990
Jeanne C. Buchodolski, Reg. No. 34,936
J. Steven Whitaker, Reg. No. 36,963

Send Correspondence to: Richard L. Hughes TOWNSEND and TOWNSEND KHOURIE and CREW Steuart Street Tower One Market Plaza, 20th Floor San Francisco, CA 94105	Direct Telephone Calls to: (Name, Reg. No., Telephone No.) Name: Richard L. Hughes Reg. No. 31,264 Telephone: (200) 467-9600

(Page 1 of 2)

Full Name of Inventor 1	Last Name YATES, III	First Name JOHN	Middle Name or Initial R
Residence & Citizenship	City Seattle	State/Foreign Country Washington WA	Country of Citizenship U.S.A.
Post Office Address	Post Office Address 8023 Forest Drive, N.E.	City Seattle	State/Country WA	Zip Code 98115
Full Name of Inventor 2	Last Name ENG	First Name JAMES	Middle Name or Initial K
Residence & Citizenship	City Seattle	State/Foreign Country Washington WA	Country of Citizenship U.S.A.
Post Office Address	Post Office Address 4860 S. 177th Ct.	City Seattle	State/Country WA	Zip Code 98128
Full Name of Inventor 3	Last Name	First Name	Middle Name or Initial
Residence & Citizenship	City	State/Foreign Country	Country of Citizenship
Post Office Address	Post Office Address	City	State/Country	Zip Code

I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Signature of Inventor 1	Signature of Inventor 2	
[signature]	*[signature]*	
John R. Yates, III	James K. Eng	
Date 3/14/94	Date 3/14/94	

solohomo 140   (Page 2 of 2)

VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY STATUS
(37 CFR 1.9(f) and 1.27(d)) - NONPROFIT ORGANIZATION

Applicant or Patentee: John R. Yates, III and James K. Eng

Serial or Patent No.: Not Assigned

Filed or Issued: Herewith

Title: USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO IDENTIFY AMINO ACID SEQUENCES IN DATABASES

I hereby declare that I am an official empowered to act on behalf of the nonprofit organization identified below:

NAME OF NONPROFIT ORGANIZATION: THE BOARD OF REGENTS OF THE UNIVERSITY OF WASHINGTON

ADDRESS OF NONPROFIT ORGANIZATION: 1107 N.E. 45th Street, Suite 200, Seattle, WA 98105

TYPE OF NONPROFIT ORGANIZATION

[x] UNIVERSITY OR OTHER INSTITUTION OF HIGHER EDUCATION
[ ] TAX EXEMPT UNDER INTERNAL REVENUE SERVICE CODE [26 USC 501(a) and 501(c)(3)]
[ ] NONPROFIT SCIENTIFIC OR EDUCATIONAL UNDER STATUTE OF STATE OR THE UNITED STATES OF AMERICA
    (NAME OF STATE   )
    (CITATION OF STATUTE   )

[ ] WOULD QUALIFY AS TAX EXEMPT UNDER INTERNAL REVENUE SERVICE CODE [26 USC 501(a) and 501(c)(3)] IF LOCATED IN THE UNITED STATES OF AMERICA
[ ] WOULD QUALIFY AS NONPROFIT SCIENTIFIC OR EDUCATIONAL UNDER STATUTE OF STATE OF THE UNITED STATES OF AMERICA IF LOCATED IN THE UNITED STATES OF AMERICA

    (NAME OF STATE   )
    (CITATION OF STATUTE   )

I hereby declare that the nonprofit organization identified above qualifies as a nonprofit organization as defined in 37 CFR 1.9(e) for purposes of paying reduced fees to the United States Patent and Trademark Office regarding the invention, entitled USE OF MASS SPECTROMETRY FRAGMENTATION PATTERNS OF PEPTIDES TO IDENTIFY AMINO ACID SEQUENCES IN DATABASES by inventor(s) John R. Yates, III and James K. Eng described in:

[x] the specification filed herewith.
[ ] application Serial No. _____ , filed _____
[ ] Patent No. _____ , issued _____

I hereby declare that rights under contract or law have been conveyed to and remain with the nonprofit organization regarding the above identified invention.

If the rights held by the nonprofit organization are not exclusive, each individual, concern or organization having rights in the invention is listed below* and no rights to the invention are held by any person, other than the inventor, who would not qualify as an independent inventor under 37 CFR 1.9(c) if that person made the invention, or by any concern that would not qualify as a small business concern under 37 CFR 1.9(d) or a nonprofit organization under 37 CFR 1.9(e).

*NOTE: Separate verified statements are required from each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)

NAME
ADDRESS _____

| [ ] INDIVIDUAL | [ ] SMALL BUSINESS CONCERN | [ ] NONPROFIT ORGANIZATION |

NAME
ADDRESS _____

| [ ] INDIVIDUAL | [ ] SMALL BUSINESS CONCERN | [ ] NONPROFIT ORGANIZATION |

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b)).

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

NAME OF PERSON SIGNING _Maryann Wenzel Toll_ margaret wenzel, Jr.

TITLE IN ORGANIZATION OF PERSON SIGNING _acting director_

ADDRESS OF PERSON SIGNING _University of Washington, JD-50, Seattle 98105_

SIGNATURE _Maryann Wenzel Toll_ DATE _3/14/94_

UNKNOWN

TANDEM MASS
SPECTROMETER

FRAGMENT SPECTRUM

PROTEIN SEQUENCE
LIBRARY

Ile Ser Glu ...
Leu Gly Leu ...
Trp Thr Ala ...
⋮
Asp Leu Arg ...
⋮

Asp Leu Arg Ser ...

AMINO ACID
SEQUENCE

COMPARE

(PRIOR ART)

FIG. 1.

UNKNOWN

TANDEM MASS
SPECTROMETER

FRAGMENT SPECTRUM

Ile Ser Glu ...

Leu Gly Leu ...

Trp Thr Ala ...

Asp Leu Arg ...

PROTEIN SEQUENCE
LIBRARY

PREDICTED MASS
SPECTRA

COMPARE

COMPUTER

FIG. 2.

FIG. 3.

INITIALIZE $m = 0$ — 36

$m = m + 1$ — 38

INITIALIZE SUM = 0 — 40
$n = 0$

$n = n + 1$ — 42

SUM = SUM + MW $(m + n)$ — 44
STORE SUM FOR USE
IN FRAGMENT CALCULATION

SUM < — 46
[MW(TARGET) – TOLERANCE]
?

YES

NO

SUM < — 48
[MW(TARGET) + TOLERANCE]
——— ?

YES

NO

MARK THE m THROUGH — 50
$m + n$
SUB-SEQUENCE AS A CANDIDATE

FIG. 4.

FIG. 5.

PREPROCESSING

Acquire data from mass spectrometer
602

Save data to file and convert to ASCII format
604

START
606

Get peptide mass and precursor ion charge state from user
608

Load ASCII mass/intensity values, rounding to unit masses
610

Remove precursor ion contribution from data
612

Normalize remaining data to maximum intensity of 100
614

Note presence of immonium ions H, F, and Y
616

Store peptide mass and immonium ion info. in datafile
618

Select top 200 most intense peaks
620

If two peaks are within 2 amu of each other, set lower intensity equal to the greater intensity
622

Store in datafile for preliminary scoring
624

Divide data into 10 windows
626

Normalize to maximum intensity of 50 within each window
628

Store in datafile for final correlation scoring
630

END
632

Fig. 6.7

DATABASE SEARCH START — *634*

Load search parameters and data from preprocessing — *636*

Load 'batch' of database sequences — *638*

Run *SEARCH* on a protein — *640*

At end of batch ? — *644*

no → Incr — *642*

yes

At end of database ? — *628*

no → Incr — *646*

yes — *630*

CORRELATION ANALYSIS

Print results — *632*

END — *634*

Fig 6B

SEARCH (START) 638

646
Start at beginning of peptide, I1 = 0

Pmass is the mass of the candidate peptide as it is being summed.

I1 is the index of the start position of the candidate peptide within the amino acid being searched.

I2 is the index of the end position of the candidate peptide within the amino acid being searched.

I2=I1, Pmass = 0   648

Add mass of amino acid at position I2 to Pmass   652

654
Does sum=(mass) input mass?   no

650

yes   656
ANALYZE candidate peptide (I1..I2)

658   660
(incr)   Is Pmass < input mass?
no

166   662
(incr)   Is I1 at end of protein?
no

yes

664
(EXIT)

1.6C

ANALYZE (START) — 670

Generate b- and y-ions for candidate peptide — 672

Is peak in top 200 ions? — 674

Sum peak intensity and increment fragment match — 676

Previous b- or y-ion matched? — 678

yes → Increment β — 680

no

All frag. ions analyzed? — 682

no → Incr — 684

yes — 686

Calculate preliminary score Sp

Is Sp > lowest score? — 688

Is sequence already stored? — 691

no → Store peptide sequence — 690

→ EXIT — 692

yes

Fig 6 D

**CORRELATION
ANALYSIS**

START

Select
stored
candidate
peptide — 693

Create theoretical
spectrum for
candidate peptide — 694

Correlate
theoretical
spectrum with
experimental
data — 695

Store final
correlation
score — 696

697

incr ← no — All
candidate
peptides
scored
? — 698

yes

EXIT — 699

Fig 6E